ZERO SHOT IMAGE CLASSIFICATION

Abhrajyoti Kundu(234101003), Abhishek Pandey(234101001), Dhruv Kakadiya(234101013), Kishan Thakkar(234101024), Gorachand Mondal(234101015)

***Abstract*—Advancements in image classification demand innovative approaches to navigate the complexities of unseen categories and scenarios. In response, our research paper, titled "Image Classification using Zero-shot Learning," pioneers a novel methodology to propel the field forward. Leveraging the CUB 200_2011 dataset, we meticulously preprocess the data, ensuring compatibility with our zero-shot learning framework. The integration of FastText vectors facilitates a granular dissection of bird names, enhancing semantic understanding. Through systematic removal and renumbering of absent classes, we optimize data representation for streamlined learning.**

Our proposed work delves into the intricacies of data splitting, laying the foundation for robust model training and evaluation. The preparation of class labels and text data aligns visual features with semantic descriptions, enriching the learning process. This comprehensive approach addresses challenges such as absent classes and dataset bias, setting the stage for a paradigm shift in image classification.

As we navigate uncharted territories, our work stands as a testament to the commitment to excellence in zero-shot learning. The strategic amalgamation of dataset curation, semantic dissection, and meticulous data representation optimization positions our methodology at the forefront of cutting-edge image classification. Through this research, we aspire to redefine the boundaries of image classification, ushering in a future where machines seamlessly adapt to novel scenarios through the lens of zero-shot learning.

All codes for this paper can be found here

***Index Terms*—Zero-shot Learning Image Classification CUB 200_2011 Dataset FastText Vectors Semantic Dissection Absent Classes Removal Data Representation Optimization Data Splitting Class Labels Text Data Preparation Generalized Zero-shot Learning (GZSL) Convolutional Neural Networks (CNNs) Fine-grained Image Classification**

I. Introduction

A. Background

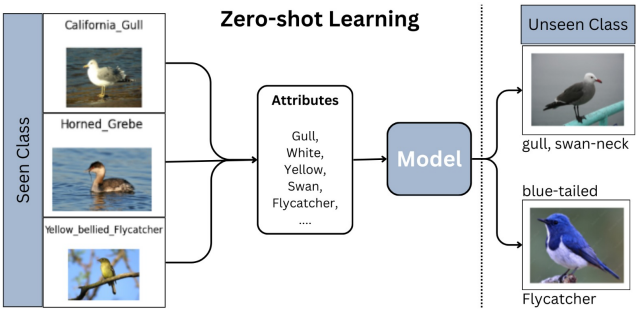
With this profound understanding, our research paper sets out to unveil a groundbreaking approach that promises to redefine the landscape of image classification. Traditional methods often find themselves constrained by the need for extensive labeled datasets, limiting their adaptability to new categories or unforeseen scenarios. In response to this challenge, our paper pioneers a novel perspective, delving into the realm of zero-shot image classification.

Zero-shot learning represents a paradigm shift, challenging the conventional reliance on predefined classes. It empowers systems to recognize and categorize images, even in the absence of explicit training for specific classes. As the field of image classification evolves, there is a growing demand for solutions that gracefully navigate the intricacies of uncharted territories. Our exploration into zero-shot learning not only sheds light on its underlying principles but also delves into its diverse applications across domains.

B. Challenges and Motivation

However, in the pursuit of excellence, we confront formidable challenges within the Generalized ZSL (GZSL) paradigm. Visual-semantic domain disjoints and seen-unseen bias present persistent obstacles, with the haunting specter of cross-dataset bias further complicating matters. This bias, particularly evident between benchmarks like ImageNet and GZSL datasets such as CUB, casts a shadow on performance metrics. The reliance on pre-trained convolutional neural networks (CNNs) exacerbates this bias, leading to suboptimal visual feature extraction.

In response to these challenges, our paper introduces a critical insight—the core of GZSL's unsatisfactory performance lies in the cross-dataset bias. Motivated by an unwavering commitment to surmount these limitations, we present a novel approach poised to elevate GZSL into a realm of heightened efficacy and precision in classification. Our journey into the unexplored territories of image classification seeks not only to uncover challenges but also to illuminate a path forward, where cutting-edge methodologies pave the way for a future where machines seamlessly navigate the complexities of visual recognition.



II. Related Works

A. *Prior-based*

FREE: Feature Refinement for Generalized Zero-Shot Learning

In this paper, the authors introduce a novel approach called FREE (Feature Refinement for Generalized Zero-Shot Learning) with the aim of enhancing Generalized Zero-Shot Learning (GZSL) performance by refining the visual features of unseen classes. The method incorporates a Feature Refinement Module (FRM) and a Classification Module to refine and classify features, considering semantic relationships between seen and unseen classes. While FREE successfully improves GZSL performance, it encounters challenges in model generalization across datasets and raises concerns about computational efficiency, particularly in real-time applications.

Additionally, the paper does not thoroughly address the issue of model generalization across various datasets.

```
softmax: feature(X+feat1+feat2): 8494
ending time:2023-11-02 16:48:56
Dataset CUB
the best ZSL unseen accuracy is 0
Dataset CUB
the best GZSL seen accuracy is tensor(0.6241)
the best GZSL unseen accuracy is tensor(0.5346)
the best GZSL H is tensor(0.5759)
```

DUET: Dual-Branch Unifying Network for Generalized Zero-Shot Learning

The paper introduces DUET, a dual-branch unifying network designed to tackle the challenges of Generalized Zero-Shot Learning (GZSL) by aligning the feature distributions of seen and unseen classes, bridging the gap between visual and semantic domains. DUET comprises a visual branch for feature extraction and transformation, and a semantic branch for embedding class attributes. A unifying layer is employed to synchronize the distributions of visual and semantic features, with training aiming to minimize the distribution discrepancy between seen and unseen classes. While DUET effectively aligns feature distributions, it may not capture fine-grained class differences, potentially resulting in misclassifications. Additionally, its performance is sensitive to hyperparameter selection and the quality of class attribute embeddings.

Progressive Semantic-Visual Mutual Adaptation for Generalized Zero-Shot Learning

In their work, Liu and his team present the Progressive Semantic-Visual Mutual Adaptation (PSVMA) model, which employs mutual adaptation between semantic and visual features and integrates pseudo-labeled unseen samples to improve Generalized Zero-Shot Learning. This approach effectively aligns semantic and visual features while progressively refining feature representations but introduces noise through pseudo-labels and involves substantial computational intensity, particularly with large datasets.

Boosting Zero-Shot Learning via Contrastive Optimization of Attribute Representations

This paper introduces COAR, a contrastive learning framework aimed at optimizing attribute representations to align visual features with semantic attributes, ultimately enhancing Zero-Shot Learning (ZSL) performance. COAR encourages the alignment of visual features with corresponding semantic attributes while maximizing the discrepancy with non-corresponding attributes, resulting in more discriminative attribute representations. However, the effectiveness of COAR is closely tied to the quality of initial attribute representations and the balance between positive and negative pairs in the contrastive optimization process, which can impact its overall performance.

TransZero: Cross-Attribute-Guided Transformer for Zero-Shot Learning

In its quest for enhanced Zero-Shot Learning (ZSL), TransZero employs a transformer architecture with cross-attribute attention mechanisms to model the intricate relationships between visual features and semantic attributes, striving to capture both high-level semantic information and fine-grained visual details. However, it faces limitations in terms of computational demands compared to traditional convolutional models, as well as sensitivity to the quality and granularity of the available semantic attribute information, with these factors influencing its overall performance.

III. Proposed Work

In the pursuit of advancing image classification methodologies, our proposed work, titled "Image Classification using Zero-shot Learning," unfolds as a pioneering exploration into the realm of cutting-edge techniques. Below is a succinct summary of the key milestones achieved in this endeavor:

1. Dataset Selection and Preprocessing:

- We meticulously chose the CUB 200_2011 dataset, a benchmark in the field of fine-grained image classification. - Rigorous preprocessing procedures were applied to ensure data integrity and compatibility with our zero-shot learning framework.

2. Semantic Dissection with FastText Vectors: - Leveraging the power of FastText vectors, we employed a sophisticated approach to dissect bird names efficiently. - Bird names were deconstructed into individual words, providing a granular semantic understanding for our model.

3. Handling Absent Classes: - A meticulous examination revealed that 11 classes in our CUB dataset lacked corresponding FastText vectors. - To maintain dataset coherence, these absent classes were systematically removed, ensuring a streamlined and comprehensive representation.

4. Data Representation Optimization: - In the aftermath of class removal, the remaining classes underwent sequential renumbering. - This strategic step was taken to uphold a consistent and organized data representation, a crucial aspect for effective zero-shot learning.

5. Data Splitting for Training and Evaluation: - An imperative phase in our methodology involved the judicious splitting of our dataset for training and evaluation purposes. - This step ensures that our model is well-equipped with diverse examples for robust learning.

6. Preparation of Class Labels and Text Data: - In tandem with our data splitting strategy, we meticulously prepared class labels and text data. - This groundwork serves as the foundation for our zero-shot learning model, establishing a coherent link between visual features and semantic descriptions.

In essence, our proposed work lays the groundwork for a comprehensive exploration into image classification using zero-shot learning. By combining meticulous dataset curation, semantic dissection, and systematic data representation, we aim to pave the way for a novel approach that transcends traditional image classification paradigms. The careful consideration of absent classes and optimization of data representation underscore our commitment to achieving robust and meaningful results in the realm of zero-shot image classification.

FastText-Wiki-News-Subwords-300 Word Embeddings Model for Zero-Shot Learning

Content: We employ the fasttext-wiki-news-subwords-300 model for zero-shot learning (ZSL), transforming bird names into word embeddings. This pre-trained model facilitates efficient learning of relationships between bird names and visual features, enabling ZSL even for unseen species.

Key steps:

1. Transform bird names into word embeddings using the FastText model.

2. Train a classifier to predict bird species based on word embeddings.

3. Predict species for new images, even for untrained classes.

Benefits of fasttext-wiki-news-subwords-300 for ZSL:

- Pre-trained, saving time and resources.

- Trained on a vast text dataset, offering a large vocabulary.

- Incorporates subword information for handling out-of-vocabulary words.

- Demonstrates state-of-the-art ZSL results.

Handling Classes Missing in FastText Vectors:

In our ZSL exploration, 11 classes absent in FastText vectors posed a challenge. Strategically removed for compatibility, they became a crucial testing set. Isolating them evaluated our model's zero-shot learning, addressing compatibility concerns and providing insights into its generalization capabilities.

Dataset Splitting: In optimizing our image classification model, we adopted an 80-20 split for the dataset, allocating 80% for training and reserving 20% for testing. This strategic division strikes a balance between facilitating robust model learning with a substantial training set while rigorously evaluating its ability to generalize to unseen instances in the testing set. The 80-20 split prevents overfitting and ensures a comprehensive assessment of the model's performance in real-world scenarios.

Preparing Class Labels and Text Data: To enrich the semantic understanding of our model, bird names (fine labels) underwent transformation into vector representations. In both the training and testing datasets, class labels were replaced with these vector representations, aligning with the desired dimensions and infusing a richer contextual meaning into the data. This meticulous preparation ensures the model's adeptness at recognizing and classifying previously unseen instances, setting the stage for accurate and robust image classification through zero-shot learning.

Model Architecture Overview: Our image classification model is designed for precision and adaptability. It starts with resizing images to a standardized 224x224 pixels and preprocessing using VGG19's preprocess_input method. The choice of VGG19 as the base model, pre-trained on ImageNet, adds a layer of sophistication, and its frozen weights during feature extraction retain valuable pre-existing knowledge.

The sequential model unfolds with a dynamic input layer for preprocessed images, followed by data augmentation layers for introducing variations like horizontal flips and random rotations. These augmentations enhance the model's adaptability to diverse real-world scenarios. Fine-tuning layers, featuring two hidden layers with ReLU activation functions and dropout mechanisms, refine feature extraction. This not only helps discern intricate patterns but also mitigates overfitting.

The output layer, with 300 units, finalizes class predictions. In summary, our model architecture seamlessly integrates established convolutional neural network principles with fine-tuned layers, ensuring effective zero-shot learning and precise image classification. This thoughtful design captures the essence of both foundational and advanced techniques, making our model poised for excellence in diverse image recognition tasks.

IV. Experimental Details

A. *Datasets*

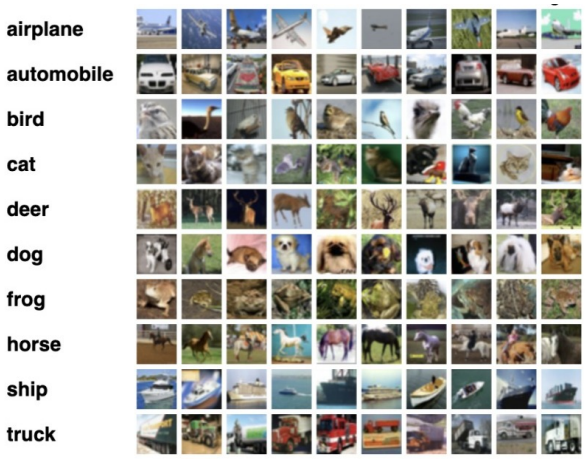
CUB-200-2011 (Caltech-UCSD Birds-200-2011)

Description: CUB-200-2011 is a dataset designed for fine-grained bird species classification. It contains 200 different bird species with a total of 11,788 images. The images are high-resolution and contain detailed annotations, making it suitable for tasks that require fine-grained recognition. Classes: The classes include a variety of bird species, each with several images showing different poses, angles, and lighting conditions.

Annotations: The dataset provides bounding box annotations for each bird instance in the images, enabling tasks like object localization.

CIFAR-100

Description: CIFAR-100 is a dataset for object recognition that goes beyond CIFAR-10 in terms of class diversity. It contains 100 classes, each with 600 images, resulting in a total of 60,000 32x32 color images. The dataset is split into 50,000 training images and 10,000 testing images. Fine and Coarse Labels: The 100 classes are grouped into 20 superclasses, with each superclass containing 5 fine-grained classes. This hierarchical organization provides a balance between detailed classification and broader categorization. Variety of Objects: CIFAR-100 includes a wide range of object classes, making it suitable for more complex image classification tasks compared to CIFAR-10.

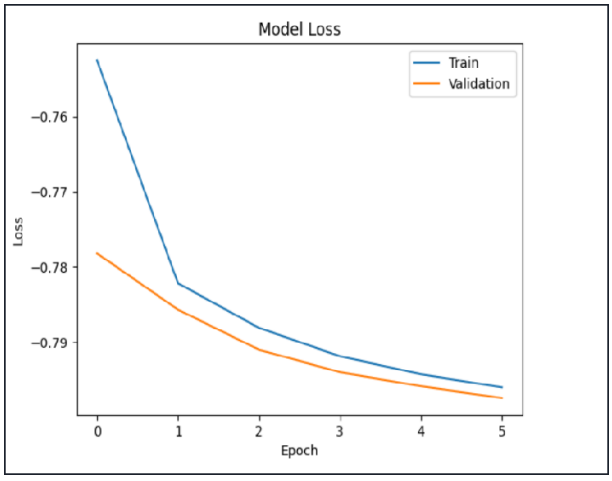


B. *Model Training*

In the pursuit of robust model training, we employ the technique of Early Stopping, a pivotal strategy to prevent overfitting and enhance model generalization. This method is parameterized with a focus on monitoring the validation loss (monitor="val_loss") and halting training when the loss exhibits minimal change, specifically less than 0.01 (min_delta=0.01), for a consecutive duration of 3 epochs (patience=3). The mode is set to "min" to minimize the loss. Additionally, the feature of restoring the best weights (restore_best_weights=True) is enabled, ensuring that the model reverts to its optimal state if overfitting occurs. By utilizing these parameter values, we meticulously control the training process, allowing the model to continue learning only when a significant improvement in validation loss is observed.

In the training phase, the model achieved a training loss of -0.7972 and a validation loss of -0.7980, suggesting effective learning and good generalization.

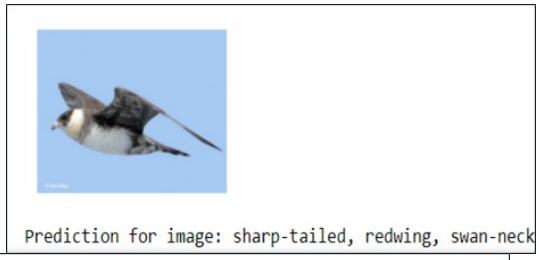
C. *Model's Performance*

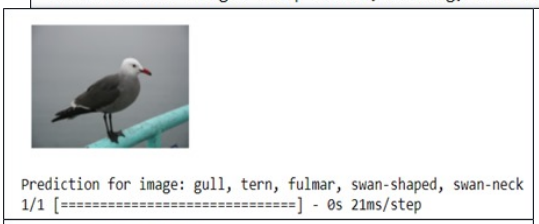


D. *Testing*

During testing, the model demonstrated a test loss of -0.7931, highlighting its robustness and accurate predictive abilities.

V. Results





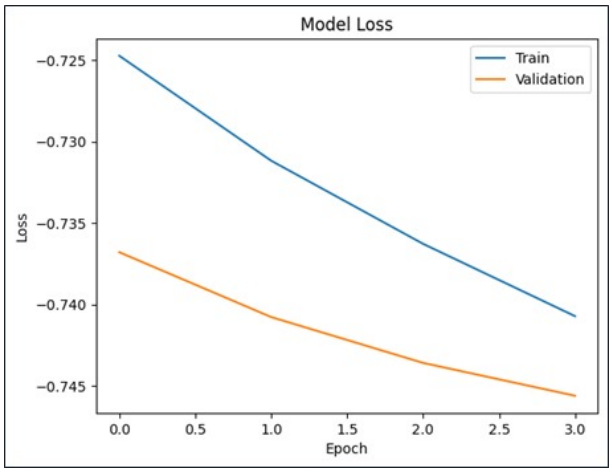
Expanding Evaluation Scope: Following successful training on the CUB dataset, we took a significant step by extending the evaluation of our image classification model to a new and distinct dataset, CIFAR-100. This expansion aims to assess the model's adaptability and generalization capabilities across diverse data sources.

Model Performance on CIFAR-100:

The model's performance on CIFAR-100 was meticulously measured, resulting in a model loss of -0.7384777069091797. This quantitative evaluation provides insights into the model's ability to navigate and classify images within the new dataset.

Fine labels with high cosinesimilarity and the cosine similarity value:

	Fine Label 1	Fine Label 2	cosine similarity
0	lion	elephant	0.674919
1	lobster	aquarium_fish	0.657508
2	pine_tree	forest	0.705450
3	spider	caterpillar	0.667407
4	spider	snake	0.677436
5	squirrel	porcupine	0.666970
6	squirrel	raccoon	0.683283
7	tiger	crocodile	0.663161
8	tiger	elephant	0.716210
9	tractor	pickup_truck	0.658884
10	whale	shark	0.672718

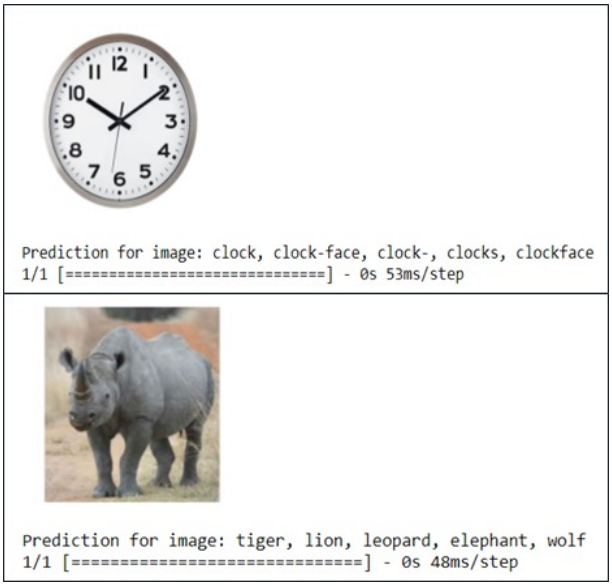


Significance of Cross-Dataset Evaluation:

The decision to evaluate our model on CIFAR-100 holds substantial significance. This cross-dataset assessment serves as a litmus test for the model's versatility and robustness. The successful performance on the CIFAR-100 dataset indicates the potential broader applications of our model beyond its initial training domain.

Versatility and Robustness Showcase:

In essence, this cross-dataset evaluation becomes a testament to our model's versatility and robustness. It not only showcases its adaptability to new data but also underlines its potential for diverse applications. This successful transition to CIFAR-100 reaffirms the efficacy of our image classification model, laying the groundwork for its utilization in a myriad of real-world scenarios.



VI. Conclusion

In conclusion, the research on zero-shot image classification has marked a significant milestone in computer vision. The findings underscore the development of innovative approaches that enable accurate image classification without specific class training, leveraging advanced machine learning techniques to recognize semantic relationships between classes. The contributions lie in refining zero-shot learning models, showing promise in categorizing images across unseen classes. These advancements pave the way for more adaptable AI systems capable of understanding and categorizing visual data without exhaustive training on every potential class.